Where was this picture painted ? – Localizing paintings by alignment to 3D models

Mathieu Aubry
INRIA*- TUM

Bryan Russell
Intel Inc.

Josef Sivic
INRIA*

mathieu.aubry@polytechnique.org

Résumé

Cet article présente une technique qui peut de manière fiable aligner une représentation non photo-réaliste d'un site architectural, tel un dessin ou une peinture, avec un model 3D du site. Pour ce faire, nous représentons le model 3D par un ensemble d'éléments discriminatifs qui sont automatiquement découverts dans des vues du modèle. Nous montrons que les éléments trouvés sont reliés de manière robuste aux changements de style (aquarelle, croquis, photographies anciennes) et aux différences structurelles. D'avantage de détails sur notre méthode et une évaluation plus détaillée est disponible [1].

Mots Clef

Reconaissance, analyse 3D , localisation.

Abstract

This paper describes a technique that can reliably align non-photorealistic depictions of an architectural site, such as drawings and paintings, with a 3D model of the site. To achieve this, we represent the 3D model by a set of discriminative visual elements that are automatically learnt from rendered views. We show that the learnt visual elements are reliably matched in 2D depictions of the scene despite large variations in rendering style (e.g. watercolor, sketch, historical photograph) and structural changes of the scene. More details and results are available in [1].

Keywords

Recognition, 3D analysis, localization.

1 Introduction

In this work we seek to automatically align historical photographs and non-photographic renderings, such as paintings and line drawings, to a 3D model of an architectural site, as illustrated in figure 1. Specifically, we wish to establish a set of point correspondences between local structures on the 3D model and their respective 2D depictions. The established correspondences will in turn allow us to find an approximate viewpoint of the 2D depiction with respect to the 3D model. We focus on depictions that are, at least approximately, perspective renderings of the 3D scene. We consider complex textured 3D models obtained by recent multi-view stereo reconstruction systems [14] as well as simplified models obtained from 3D modeling tools such as Trimble 3D Warehouse.

FIGURE 1 – Our system automatically recovers the viewpoint of paintings, drawings, and historical photographs by aligning the input painting (left) with the 3D model (right).

This task is extremely challenging. As discussed in prior work [27, 31], local feature matching based on interest points (e.g. SIFT [23]) often fails to find correspondences across paintings and photographs. First, the rendering styles across the two domains can vary considerably. The scene appearance and geometry depicted by the artist can be very different from the rendering of the 3D model, e.g. due to the depiction style or drawing error. Second, we face a hard search problem. The number of possible alignments of the painting to a large 3D model, such as a partial reconstruction of a city, is huge. Which parts of the painting should be aligned to which parts of the 3D model ? How to search over the possible alignments ?

To address these issues we introduce the idea of automatically discovering *discriminative visual elements* for a 3D scene. We define a discriminative visual element to be a mid-level patch that is rendered with respect to a given viewpoint from a 3D model with the following properties : (i) it is visually discriminative with respect to the rest of the "visual world" represented here by a generic set of randomly sampled patches, (ii) it is distinctive with respect to other patches in nearby views, and (iii) it can be reliably matched across nearby viewpoints. We employ modern representations and recent methods for discriminative lear-

*WILLOW project-team, Département d'Informatique de l'Ecole Normale Supérieure, ENS/INRIA/CNRS UMR 8548

ning of visual appearance, which have been successfully used in recent object recognition systems. Our method can be viewed as “multi-view geometry [18] meets part-based object recognition [11]”.

We discover discriminative visual elements by first sampling candidate mid-level patches across different rendered views of the 3D model. We cast the image matching problem as a classification task over appearance features with the candidate mid-level patch as a single positive example and a negative set consisting of a large set of “background" patches. Note that a similar idea has been used in learning per-exemplar distances [13] or per-exemplar support vector machine (SVM) classifiers [25] for object recognition and cross-domain image retrieval [31].

For a candidate mid-level patch to be considered a discriminative visual element, we require that (i) it has a low training error when learning the matching classifier, and (ii) it is reliably detectable in nearby views via cross-validation. Critical to the success of operationalizing the above procedure is the ability to efficiently train linear classifiers over Histogram of Oriented Gradients (HOG) features [7] for each candidate mid-level patch, which has potentially millions of negative training examples. In contrast to training a separate SVM classifier for each mid-level patch, we change the loss to a square loss, similar to [4, 16], and show that the solution can be computed in closed-form, which is computationally more efficient as it does not require expensive iterative training. In turn, we show that efficient training opens-up the possibility to evaluate the discriminability of millions of candidate visual elements densely sampled over all the rendered views. We show that our approach is able to scale to a number of different 3D sites and handles different input rendering styles. Moreover, we are able to handle different types of 3D models, such as 3D CAD models and models constructed using multi-view stereo [15]. To evaluate our alignment procedure, we introduce a database of paintings and sketches spanning several sites and perform a user study where human subjects are asked to judge the goodness of the output alignments. Moreover, we evaluate our matching step on the benchmark dataset of [19] and show improvement over local symmetry features [19] and several alternative matching criteria for our system.

2 Prior work

Alignment. Local invariant features and descriptors such as SIFT [23] represent a powerful tool for matching photographs of the same at least lightly textured scene despite changes in viewpoint, scale, illumination, and partial occlusion. Large 3D scenes, such as a portion of a city [22], can be represented as a 3D point cloud with associated local feature descriptors extracted from the corresponding photographs [28]. Camera pose of a given query photograph can be recovered from 2D to 3D correspondences obtained by matching appearance of local features verified using geometric constraints [18]. However, appearance changes beyond the modeled invariance, such as significant perspective distortions, non-rigid deformations, non-linear illumination changes (e.g. shadows), weathering, change of seasons, structural variations or a different depiction style (photograph, painting, sketch, drawing) cause local feature-based methods to fail [19, 27, 31]. Greater insensitivity to appearance variation can be achieved by matching the geometric or symmetry pattern of local image features [6, 19, 30], rather than the local features themselves. However, such patterns have to be detectable and consistent between the matched views. An alternative to feature-based alignment is contour based alignment [20, 24]. Recent work [2, 3] has shown that it is a powerful tool when contours as skyline can be accurately extracted. However, that is rarely the case, especially for paintings and real world 3D meshes.

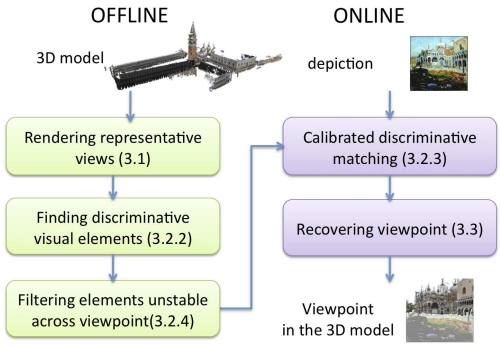

FIGURE 2 – **Approach overview.** In the offline stage (left) we summarize a given 3D model as a collection of discriminative visual elements. In the online stage (right) we match the learnt visual elements to the input painting and use the obtained correspondences to recover the camera viewpoint with respect to the 3D model.

Discriminative learning. Modern image representations developed for visual recognition, such as HOG descriptors [7], represent 2D views of objects or object parts [11] by a weighted spatial distribution of image gradient orientations. The weights are learnt in a discriminative fashion to emphasize object contours and de-emphasize non-object, background contours and clutter. Such a representation can capture complex object boundaries in a soft manner, avoiding hard decisions about the presence and connectivity of imaged object edges. Learnt weights have also been shown to emphasize visually salient image structures matchable across different image domains, such as sketches and photographs [31]. Similar representation has been used to learn architectural elements that summarize a certain geo-spatial area by analyzing (approximately rectified) 2D street-view photographs from multiple cities [9].

3 Approach overview

The proposed method has two stages : first, in an offline stage we learn a set of discriminative visual elements representing the architectural site ; second, in an online stage a given unseen query painting is aligned with the 3D model by matching with the learnt visual elements. The proposed

algorithm is summarized in figure 2.

3.1 Rendering representative views

We sample possible views of the 3D model in a similar manner to [2, 21, 27]. First, we identify the ground plane and corresponding vertical direction. The camera positions are then sampled on the ground plane on a regular grid. For each camera position we sample 12 possible horizontal camera rotations assuming no in-plane rotation of the camera. For each horizontal rotation we sample 2 vertical rotations (pitch angles). Views where less than 5% of the pixels are occupied by the 3D model are discarded. This procedure results in 7,000-45,000 views depending on the size of the 3D site. Note that the rendered views form only an intermediate representation and can be discarded after visual element detectors are extracted.

3.2 Discriminative visual elements

Matching as classification. The aim is to match a given rectangular image patch q (represented by a HOG descriptor [7]) in a rendered view to its corresponding image patch in the painting, as illustrated in figure 3. Instead of finding the best match measured by the Euclidean distance between the descriptors, we train a linear classifier with q as a single positive example (with label $y_q = +1$) and a large number of negative examples x_i for $i = 1$ to N (with labels $y_i = -1$). The matching is then performed by finding the patch x^* in the painting with the highest classification score

$$s(x) = w^\top x + b, \tag{1}$$

where w and b are the parameters of the linear classifier. Parameters w and b can be obtained by minimizing a cost function of the following form

$$E(w, b) = L\left(1, w^T q + b\right) + \frac{1}{N}\sum_{i=1}^{N} L\left(-1, w^T x_i + b\right), \tag{2}$$

where the first term measures the loss L on the positive example q (also called “exemplar") and the second term measures the loss on the negative data. A particular case of the exemplar based classifier is the exemplar-SVM [25, 31], where the loss $L(y, s(x))$ between the label y and predicted score $s(x)$ is the hinge-loss $L(y, s(x)) = \max\{0, 1 - ys(x)\}$ [5]. For exemplar-SVM cost (2) is convex and can be minimized using iterative algorithms [10, 29], but this remains computationally expensive.

Selection of discriminative visual elements via least squares regression. Using instead a square loss $L(y, s(x)) = (y - s(x))^2$, similarly to [4, 16], w_{LS} and b_{LS} minimizing (2) and the optimal cost E^*_{LS} can be obtained in closed form as

$$w_{LS} = \frac{2}{2 + \|\Phi(q)\|^2}\Sigma^{-1}(q - \mu), \tag{3}$$

$$b_{LS} = -\frac{1}{2}(q + \mu)^T w_{LS}, \tag{4}$$

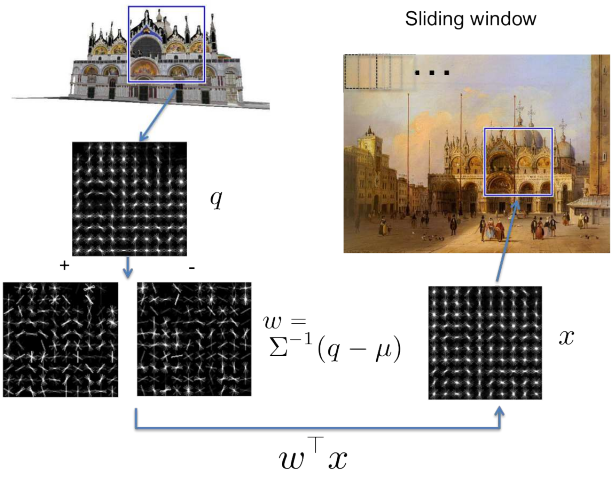

FIGURE 3 – **Matching as classification.** Given a region and its HOG descriptor q in a rendered view (top left) the aim is to find the corresponding region in a painting (top right). This is achieved by training a linear HOG-based sliding window classifier using q as a single positive example and a large number of negative data. The classifier weight vector w is visualized by separately showing the positive (+) and negative (-) weights at different orientations and spatial locations. The best match x in the painting is found as the maximum of the classification score.

$$E^*_{LS} = \frac{4}{2 + \|\Phi(q)\|^2}, \tag{5}$$

where $\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$ denotes the mean of the negative examples, $\Sigma = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)(x_i - \mu)^\top$ their covariance and Φ is the “whitening" transformation :

$$\|\Phi(x)\|^2 = (x - \mu)^\top \Sigma^{-1}(x - \mu), \tag{6}$$

We can use the value of the optimal cost (5) as a measure of the discriminability of a specific q. If the training cost (error) for a specific candidate visual element q is small the element is discriminative. This observation can be translated into a simple and efficient algorithm for ranking candidate element detectors based on their discriminability. Given a rendered view, we consider as candidates visual element all patches that are local minima (in scale and space) of the training cost 5.

Relation to linear discriminant analysis (LDA). An alternative way to compute w and b is to use LDA, similarly to [16, 17]. It results in the parameters :

$$w_{LDA} = \Sigma^{-1}(q - \mu_n), \tag{7}$$

and

$$b_{LDA} = \frac{1}{2}\left(\mu^T \Sigma^{-1}\mu - q^T \Sigma^{-1} q\right). \tag{8}$$

Nothe that w_{LDA} is proportional to w_{LS}. It implies that both method lead to the same matches.

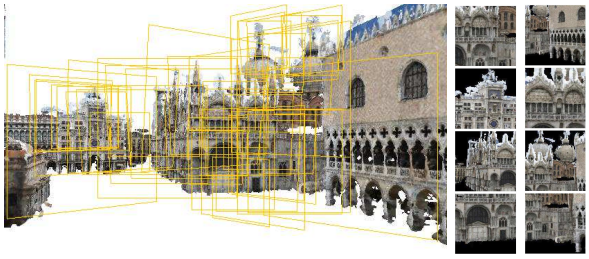

FIGURE 4 – **Examples of selected visual elements for a 3D site. Left :** Selection of top ranked 50 visual elements visible from this specific view of the site. Each element is depicted as a planar patch with an orientation of the plane parallel to the camera plane of its corresponding source view. **Right :** Subset of 8 elements shown from their original viewpoints. Note that the proposed algorithm prefers visually salient scene structures such as the two towers in the top-right or the building in the left part of the view.

Calibrated discriminative matching. We have found that calibration of matching scores across different visual elements is important for the quality of the final matching results. Below we describe a procedure to calibrate matching scores without the need of any labelled data. First, we found (section 4) that the matching score obtained from LDA produces significantly better matching results than matching via least squares regression. Nevertheless, we found that the raw uncalibrated LDA score favors low-contrast image regions, which have an almost zero HOG descriptor. To avoid this problem, we further calibrate the LDA score by subtracting a term that measures the score of the visual element q matched to a low-contrast region, represented by zero (empty) HOG vector

$$s_{calib}(x) = s_{LDA}(x) - s_{LDA}(0) \quad (9)$$
$$= (q-\mu)^T \Sigma^{-1} x. \quad (10)$$

This calibrated score gives much better results on the dataset of [19] as shown in section 4 and significantly improves matching results.

Filtering elements unstable across viewpoint. To avoid ambiguous elements, we perform two additional tests on the visual elements. First, to suppress potential repeated structures, we require that the ratio between the score of the first and second highest scoring detection in the image is larger than a threshold of 1.04, similar to [23]. Second, we run the discriminative elements in the views near the one where they were defined and keep visual elements that are successfully detected in more than 80% of the nearby views. This procedure typically results in several thousand selected elements for each architectural site. Examples of the final visual elements obtained by the proposed approach are shown in figure 4.

3.3 Recovering viewpoint

Following the matching procedure described in section 3.2, we form a set of matches using the following procedure.

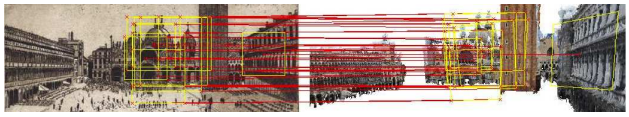

FIGURE 5 – **Illustration of alignment.** We use the recovered discriminative visual elements to find correspondences between the input scene depiction (left) and 3D model (right). Shown is the recovered viewpoint and inlier visual elements found via RANSAC.

First, we apply all visual element detectors on the depiction and take the top 200 detections sorted according to the first to second nearest neighbor ratio [23], using the calibrated similarity score (9). This selects the most non-ambiguous matches. Second, we sort the 200 matches directly by score (9) and consider the top 25 matches. From each putative visual element match we obtain 5 putative point correspondences by taking the 2D/3D locations of the patch center and its four corners. The patch corners provide information about the patch scale and the planar location on the 3D model, and has been shown to work well for structure-from-motion with planar constraints [32]. We use RANSAC [12] to find the set of inlier correspondences to a restricted camera model where the camera intrinsics are fixed, with the focal length set to the image diagonal length and the principal point set to the center of the image. The recovered viewpoint provides an alignment of the input depiction to the 3D model, which is shown in figure 5.

4 Results and validation

To evaluate our method, we have collected a set of human-generated 3D models from Trimble 3D Warehouse for the following architectural landmarks : Notre Dame of Paris, Trevi Fountain, and San Marco's Basilica. The Trimble 3D Warehouse models for these sites consist of basic primitive shapes and have a composite texture from a set of images. We also consider a 3D models of San Marco's Square that was reconstructed from a set of photographs using dense multi-view stereo [14]. Note that while the latter 3D model has more accurate geometry than the Trimble 3D Warehouse models, it is also much noisier.
We have also collected from the Internet 85 historical photographs and 252 non-photographic depictions of the sites. Figures 6 shows examples of alignment results. Notice that the depictions are reasonably well-aligned, with regions on the 3D model rendered onto the corresponding location for a given depiction. We are able to cope with a variety of viewpoints with respect to the 3D model as well as different depiction styles, challenging appearance changes and the varying quality of the 3D models.

Quantitative evaluation. To quantitatively evaluate the goodness of our alignments, we have conducted a user study via Amazon Mechanical Turk. The workers were asked to judge the viewpoint similarity of the resulting alignments to their corresponding input depictions by categorizing the

FIGURE 6 – Example alignments of non-photographic depictions to 3D models. Notice that we are able to align depictions rendered in different styles and having a variety of viewpoints with respect to the 3D models. **More results are available at the project website `http://www.di.ens.fr/willow/research/painting_to_3d/`**

TABLE 1 – Viewpoint similarity user study of our algorithm across different depiction styles.

	Good match	Coarse match	No match
Historical photographs	59%	20%	21%
Paintings	53%	30%	18%
Drawings	52%	29%	19%
Engravings	57%	26%	17%
Average	**55%**	**27%**	**18%**

TABLE 2 – **Evaluation of visual element matching.** We report the mean average precision on the "desceval" task from the benchmark dataset of [19].

Matching method	mAP ("desceval")
Local symmetry [19]	0.58
Least squares regression (Sec. 3.2)	0.52
LDA (Sec. 3.2)	0.60
Ours (Sec. 3.2)	**0.77**

viewpoint similarity as either a (a) Good match, (b) Coarse match, or (c) No match. We report the majority opinion. Table 1 shows the performance of our algorithm for different depiction styles averaged across the 3D sites. Interestingly, the results are fairly consistent across different depiction styles and the failure rate (no match) remains consistently below 25%.

Visual element matching. We evaluate the proposed matching procedure on the 'desceval' task from the benchmark dataset collected in [19]. Challenging pairs of images in the dataset depicting a similar viewpoint of the same landmark have been manually registered using a homography. The task is to find corresponding patches in each image pair. Following [19] we perform matching over a grid of points in the two views, with the grid having 25 pixel spacing. Since the ground truth correspondence between points is known, a precision-recall curve can be computed for each image pair. We report the mean average precision (mAP) measured over all image pairs in the dataset in table 2. Our full system using the calibrated matching score (section 3.2) achieves a mAP of 0.77, which significantly outperforms both the alternative visual element matching scores obtained by least squares regression (section 3.2) and linear discriminant analysis (LDA, section 3.2), as well as the local symmetry feature baseline.

5 Conclusion

We have demonstrated that automatic image to 3D model alignment is possible for a range of non-photographic depictions and historical photographs, which represent extremely challenging cases for current local feature matching methods. To achieve this we have developed an approach to compactly represent a 3D model of an architectural site by a set of visually distinct mid-level scene elements extracted from rendered views. This work is just a step towards computational reasoning about the content of non-photographic depictions.

Acknowldegments

We are grateful to Alyosha Efros, Guillaume Obozinski and Jean Ponce for their useful feedback, to Yasutaka Furukawa for providing access to the San Marco 3D model, and to Daniel Hauagge and Noah Snavely for providing the evaluation code for their paper [19]. This work was partly supported by the ANR project SEMAPOLIS, EIT ICT Labs, the MSR-INRIA laboratory and Google.

Références

[1] M. Aubry, B. Russell, and J. Sivic. Painting-to-3D model alignment via discriminative visual elements. In *ACM Trans. Graph.*, 2014. To appear.

[2] G. Baatz, O. Saurer, K. Köser, and M. Pollefeys. Large scale visual geo-localization of images in mountainous terrain. In *ECCV*, 2012.

[3] L. Baboud, M. Cadik, E. Eisemann, and H.-P. Seidel. Automatic photo-to-terrain alignment for the annotation of mountain pictures. In *CVPR*, 2011.

[4] F. Bach and Z. Harchaoui. Diffrac : a discriminative and flexible framework for clustering. In *NIPS*, 2008.

[5] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[6] O. Chum and J. Matas. Geometric hashing with local affine frames. In *CVPR*, 2006.

[7] N. Dalal and B. Triggs. Histograms of Oriented Gradients for Human Detection. In *CVPR*, 2005.

[8] T. Dean, M. Ruzon, M. Segal, J. Shlens, S. Vijayanarasimhan, and J. Yagnik. Fast, accurate detection of 100,000 object classes on a single machine. In *CVPR*, 2013.

[9] C. Doersch, S. Singh, A. Gupta, J. Sivic, and A. A. Efros. What makes paris look like paris ? *SIGGRAPH*, 31(4), 2012.

[10] R. Fan, K. Chang, C. Hsieh, X. Wang, and C. Lin. Liblinear : A library for large linear classification. *Journal of Machine Learning Research*, 9(1) :1871–1874, 2008.

[11] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part based models. *IEEE PAMI*, 32(9) :1627–1645, 2010.

[12] M. Fischler and R. Bolles. Random sample consensus : A paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6) :381–395, 1981.

[13] A. Frome, Y. Singer, F. Sha, and J. Malik. Learning globally-consistent local distance functions for shape-based image retrieval and classification. In *ICCV*, 2007.

[14] Y. Furukawa, B. Curless, S. M. Seitz, and R. Szeliski. Towards internet-scale multi-view stereo. In *CVPR*, 2010.

[15] Y. Furukawa and J. Ponce. Accurate, dense, and robust multi-view stereopsis. *IEEE PAMI*, 32(8), 2010.

[16] M. Gharbi, T. Malisiewicz, S. Paris, and F. Durand. A Gaussian approximation of feature space for fast image similarity. Technical Report MIT-CSAIL-TR-2012-032, 2012.

[17] B. Hariharan, J. Malik, and D. Ramanan. Discriminative decorrelation for clustering and classification. In *ECCV*, 2012.

[18] R. I. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, ISBN : 0521540518, second edition, 2004.

[19] D.C. Hauagge and N. Snavely. Image matching using local symmetry features. In *CVPR*, 2012.

[20] D. P. Huttenlocher and S. Ullman. Object recognition using alignment. In *International Conference on Computer Vision*, 1987.

[21] A. Irschara, C. Zach, J.-M. Frahm, and H. Bischof. From structure-from-motion point clouds to fast location recognition. In *CVPR*, 2009.

[22] Y. Li, N. Snavely, D. Huttenlocher, and P. Fua. Worldwide pose estimation using 3D point clouds. In *ECCV*, 2012.

[23] D. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2) :91–110, 2004.

[24] D. G. Lowe. Three-dimensional object recognition from single two-dimensional images. *Artificial Intelligence*, 31(3) :355–395, 1987.

[25] Tomasz Malisiewicz, Abhinav Gupta, and Alexei A. Efros. Ensemble of exemplar-svms for object detection and beyond. In *ICCV*, 2011.

[26] J. B. Rapp. A geometrical analysis of multiple viewpoint perspective in the work of Giovanni Battista Piranesi : an application of geometric restitution of perspective. *The Journal of Architecture*, 13(6), 2008.

[27] B. C. Russell, J. Sivic, J. Ponce, and H. Dessales. Automatic alignment of paintings and photographs depicting a 3D scene. In *IEEE Workshop on 3D Representation for Recognition (3dRR-11), associated with ICCV*, 2011.

[28] T. Sattler, B. Leibe, and L. Kobbelt. Fast image-based localization using direct 2D-to-3D matching. In *ICCV*, 2011.

[29] S. Shalev-Shwartz, Y. Singer, N. Srebro, and A. Cotter. Pegasos : Primal Estimated sub-GrAdient SOlver for SVM. *Mathematical Programming, Series B*, 127(1) :3–30, 2011.

[30] E. Shechtman and M. Irani. Matching local self-similarities across images and videos. In *CVPR*, 2007.

[31] A. Shrivastava, T. Malisiewicz, A. Gupta, and A. A. Efros. Data-driven visual similarity for cross-domain image matching. In *SIGGRAPH Asia*, 2011.

[32] R. Szeliski and P. Torr. Geometrically constrained structure from motion : Points on planes. In *European Workshop on 3D Structure from Multiple Images of Large-Scale Environments (SMILE)*, 1998.